图书情报工作
第65卷 第14期 2021年7月

基于事件本体的疫情知识库构建策略*

■ 熊励[1] 王成文[1] 王锟[1,2]
[1] 上海大学管理学院 上海 200444 [2]悉尼科技大学澳大利亚人工智能研究所 悉尼 2007

摘 要：[目的/意义]疫情信息碎片化和非结构化给应急决策带来了挑战。为支撑应急决策数字化和促进应急管理智能化，结合自然语言处理、事件本体实现疫情信息管理和知识表示的自动化。[方法/过程]提出一种基于网络爬虫、自然语言处理、事件本体的领域本体知识库自动构建策略。首先，运用网络爬虫和自然语言处理进行信息采集和事件要素自动提取，在此基础上构建疫情事件本体模型。然后，设计本体构建与更新算法，通过该算法完成事件本体的自动构建与扩充。[结果/结论]研究结果表明，该策略具备疫情信息动态管理与自动更新的可行性，且事件本体能够有效描述事件，并为知识的拓展创造条件。本研究为应急管理决策的相关研究与实践提供一定的参考。

关键词：事件本体 疫情信息与知识 知识库
分类号：G250
DOI：10.13266/j.issn.0252-3116.2021.14.016

疫情信息是突发事件应急决策的关键依据，大数据环境下产生的海量疫情数据和信息为疫情有效防控和应急管理能力提升提供支撑，基于新一代信息技术的突发事件信息管理方案成为学术界和相关部门的关注焦点。在新冠肺炎疫情防控中，由于疫情相关信息的碎片化、信息更新不及时、信息壁垒、情报分析能力不足等造成疫情信息共享与管理困难[1-2]，对防疫决策支持不足，最终影响了应急响应效率，制约了疫情防控的效果，造成应急管理统筹协调不力。疫情各阶段涌现的大量非结构化数据对传统的以关系数据库为载体的结构化数据采集与加工方式带来了挑战。因此，当前亟需疫情信息的动态采集与处理、疫情知识自动表示与提取、信息共享的智能化手段，在此基础上，为应急决策提供关键支撑。

在社会对公共卫生突发事件应急管理智能化的迫切需求下，笔者聚焦疫情事件，以网络疫情信息和开放知识图谱为数据源，采用文本分析和事件本体等方法和技术研究疫情信息和知识的自动表示，为疫情应急管理提供智能决策支持。

1 相关研究综述

1.1 疫情信息管理与知识表示的研究

包括疫情在内的突发事件的信息管理与知识表示是应急管理的重要内容，能够为应急管理提供决策依据[3]，而构建知识库是实现疫情信息有效获取、组织、分析和传递的重要手段[4]。通过文献梳理发现，突发事件的信息管理和知识表示主要基于信息和情报系统建模、语义本体等方案，为应急响应中的信息检索和决策提供参考和借鉴。其中，W. B. Lee等提出了一种用于应急管理的非结构化信息管理系统[5]，通过应急事件概念关系模型和动态知识流模型来组织和表示突发事件知识；M. Dorasamy等通过应急管理信息系统来解决数据管理、知识共享和传播问题[6]，为应急管理人员提供关键数据、信息和知识，促进救生信息、知识共享；郭骅等采用信息资源规划方法组织突发事件应急管理情报流，构建了城市应急管理情报平台[7]，为应急管理提供动态信息和知识。

在将本体应用于应急知识表示的研究中，突发事件网络舆情知识表示[8]、基于时空角度的突发事件社

* 本文系国家社会科学基金国家应急管理体系建设研究专项项目“基于人工智能机器学习和区块链技术支撑的疫情监测防控研究”（项目编号：20VYJ064）研究成果之一。
作者简介：熊励（ORCID：0000-0002-6527-0517），教授，博士，博士生导师；王成文（ORCID：0000-0002-7353-472X），博士研究生，通讯作者，E-mail：woncwen@163.com；王锟（ORCID：0000-0003-2711-6233），博士研究生。
收稿日期：2021-01-06 修回日期：2021-02-28 本文起止页码：138-148 本文责任编辑：徐健

会感知知识本体模型[9]、面向火灾的本体应急知识库[10]、地铁事故本体模型[11]，为突发事件的应急响应提供了决策依据。相关研究适应现代应急管理的要求，逐步从静态的应急信息管理到动态的知识表示，为应急管理提供决策支持。

疫情信息管理策略的研究相对欠缺，且主要侧重于传染病具体症状与病例信息的分析与管理。高珊等设计的传染病应急案例处置本体模型[12]，对传染病传播与处置进行知识建模；方安等面向传染病病症与诊疗的知识服务平台[13]，将概念与对象及其之间的关系映射为知识网络；W. R. Hogan 等开发了一个流行病学本体[14]，将传染病信息概念化；陈晓慧等通过本体实现了新冠肺炎病例活动知识图谱[15]，以支持传播过程和病例轨迹的分析；A. Joshi 等从健康状况、传播数量与方式、地点、时间几个维度限定传染病的概念范围[16]，为疫情知识表示提供了参考。这些研究主要聚焦于病理与诊疗、传播等角度的疫情知识组织，有助于疫情的监测和诊断，而从疫情事件角度，针对疫情整体动态演化的知识表示并为防疫决策提供支持的策略与方案的研究鲜有。因此，结合新冠肺炎疫情防控中暴露的信息效用不足的短板，笔者面向疫情辅助决策视角，以事件为脉络，探究基于动态更新的疫情信息管理和知识表示策略。

1.2 事件本体的相关研究

本体是领域内实体及其属性、实体间关系的概念化表示[17]。从基本结构来看，本体包含某一领域的基本概念(类)集、体现概念间关系的对象属性集、界定概念特征的数据属性集、表示现实世界中概念模型的实例等要素[18]，相关研究主要基于这些要素进行本体开发与应用。事件是指在特定时间和位置发生的一件事情，涉及多个参与者[19－20]，并显示了某些动作特征，具体包含动作、参与主体、事件客体、时间、地点等要素。事件本体模型是应用于发生事件的认知表示的知识架构[21]，事件本体基于该框架实现事件类知识的形式化说明和共享[22]，围绕事件主题，能够进行知识表示、语义化和推理。目前，针对事件本体的研究主要集中在通过本体建模提取和分享领域知识(事件本体的应用)、本体的建模与构建策略(事件本体的开发)两个方面。

在事件本体的应用方面，事件本体为突发事件和社会热点的知识建模提供了参考方案。其中，基于共享词汇的环境污染事件本体模型用于提取多种污染事件中的语义关系[23]；基于多源数据的洪水应急决策支持系统采用事件本体对复杂情景建模[24]；安全事故知识事件本体模型实现了安全事故案例及其场景的动态表示[25－26]，为事故分析和预测提供决策依据；基于事件本体的 Web 服务组合[27]、体育赛事的知识表示与分类[28]、新闻推荐[29]的研究为供需信息表示和兴趣分析提供了路径。这些研究丰富了事件本体的应用，为领域知识表示与辅助决策提供了参考与借鉴。

在本体建模与开发方面，人工方式仍是主要途径，但这一策略费时费力且存在手工失误的风险，提高了本体研究与应用的门槛。为此，刘思含等提出了自然语言处理与神经网络结合的事件本体自动构建的理论方案[30]，为相关研究提供了理论指导。在具体开发层面，朱文跃等围绕事件特征与结构，从事件类别、事件关系与实例两个层次来设计本体模型[31]，为事件本体建模提供了基本框架；J. I. Single 等运用自然语言处理从化学事故数据库中提取事件要素并进行事件本体填充[26]，但并未能实现实例关系的自动构建；O. Gurbuz 等通过句子主谓宾成分的提取来确定组织流程的关系与状态，并根据词性确定事件对象和状态等要素[32]，其效果依赖于文本分析的效果；J. A. Reyes-Ortiz 通过动词、名词、介词短语和词缀成分识别事件要素[33]，但对专有名词、名词短语、量词的识别和提取效果不佳；王思丽等通过词性标注对领域概念进行自动识别[34]，为本体自动构建的前期工作提供了思路；Q. Mao 等通过词性和句法分析识别事件及其要素[35]，为基于语义的事件演化分析提供形式化知识。这些研究为事件要素的识别和提取提供了参考和借鉴，但在本体构建环节多采用人工方式，难以满足大规模本体构建的要求。

综上所述，目前针对疫情信息管理与知识表示的研究仍然比较欠缺。事件本体围绕事件主题，从涉事主体、时间、地点等角度组织事件知识，为动态知识表示、事件演化分析提供了可行方案，也为疫情知识表示提供了思路。文章从疫情信息和知识的有效组织视角对疫情知识表示策略展开研究，设计面向网络数据的知识自动表示与更新方案，以强化疫情信息和知识对应急管理决策的支撑作用。

2 疫情事件本体模型构建

事件本体模型是事件的知识框架，能够刻画事件及其要素之间的关系，是事件本体的基本框架。新冠肺炎疫情产生了大量的疫情动态信息，为疫情分析和知识提取提供了信息资源，同时也增加了应急决策的

难度。笔者将从网络爬取的疫情播报作为基础数据源,参考文献[16,31]的事件类框架、文献[34－35]的领域概念与事件识别方法,采用词性分析、命名实体识别和语义角色标注提取疫情事件关键要素,围绕疫情事件要素展开事件本体模型的设计与本体构建,并在此基础上进行本体的拓展。

2.1 疫情事件要素分析

疫情事件本体的构建基于事件本体模型和动态、客观的疫情事件信息。新冠疫情防控中,各省市都积极对本辖区内的疫情事件进行了客观和及时的播报,为疫情情报的自动、及时采集提供了便利。笔者以官方疫情播报和开放知识图谱为数据源,采用爬虫技术定时采集疫情信息,运用中文自然语言处理的代表性工具 LTP 来处理疫情信息文本、规范疫情情报语料库,设计事件要素列表构建算法将语义角色标注与文本块一一对应。语义角色标注是以文本的谓词为核心,识别其他成分与谓词的关系,进而实现关键信息的提取,笔者将其作为确定疫情事件要素的依据。语义角色标注识别了疫情事件的基本要素,为疫情事件本体模型和事件本体的自动构建提供依据。表 1 展示了从上海市卫生健康委员会(简称卫健委)的疫情播报信息中提取疫情事件要素的过程。

表 1 疫情事件要素提取过程示例

文本分析	处理结果
文本读取	12 月 31 日,上海市新增 5 例境外输入性确诊病例。新增治愈出院 7 例,其中来自俄罗斯 2 例,来自巴基斯坦 1 例,来自加拿大 1 例,来自安哥拉 1 例,来自西班牙 1 例,来自斯洛文尼亚 1 例
文本分词(seg): Segs	['2020 年', '12 月', '31 日', ',', '上海市', '新增', '5', '例', '境外输入性病例', '。', '新增', '治愈出院病例', '7', '例', ',', '其中', '来自', '俄罗斯', '2', '例', ',', '来自', '巴基斯坦', '1', '例', ',', '来自', '加拿大', '1', '例', ',', '来自', '安哥拉', '1', '例', ',', '来自', '西班牙', '1', '例', ',', '来自', '斯洛文尼亚', '1', '例', '。']
命名实体识别(ner): Ns	[('ns', 4, 4), ('ns', 17, 17), ('ns', 22, 22), ('ns', 27, 27), ('ns', 32, 32), ('ns', 37, 37), ('ns', 42, 42)] [上海市,俄罗斯,巴基斯坦,加拿大,安哥拉,西班牙,斯洛文尼亚]
词性分析(pos): Pos	[['nt', 'nt', 'nt', 'wp', 'ns', 'v', 'm', 'q', 'nl', 'wp'], ['v', 'v', 'm', 'n', 'wp', 'r', 'v', 'ns', 'm', 'n', 'wp', 'v', 'ns', 'm', 'n', 'wp', 'v', 'ns', 'm', 'n', 'wp', 'v', 'ns', 'm', 'n', 'wp', 'v', 'ns', 'm', 'n', 'wp', 'v', 'ns', 'm', 'n', 'wp']]→ [[['tim', 'as', 'act', 'num', 'eo'], ['2020 年 12 月 31 日', '上海市', '新增', '5', '境外输入性病例']], [['as', 'act', 'eo', 'num', 'fro', 'so', 'fro', 'so', 'fro', 'so', 'fro', 'so', 'fro', 'so', 'fro', 'so'], ['上海市', '新增', '治愈出院病例', '7', '来自', '俄罗斯', '来自', '巴基斯坦', '来自', '加拿大', '来自', '安哥拉', '来自', '西班牙', '来自', '斯洛文尼亚']]]
语义角色标注(srl): Srl	[(5, [('TMP', 0, 2), ('A0', 4, 4), ('A1', 8, 8), ('QTY', 6, 6)]), (10, [('TMP', 0, 2), ('A0', 4, 4), ('A1', 11, 11), ('QTY', 12, 12)]), (16, [('A1', 17, 17), ('A0', 11, 11)]), (21, [('A1', 22, 22), ('A0', 11, 11)]), (26, [('A1', 27, 27), ('A0', 11, 11)]), (31, [('A1', 32, 32), ('A0', 11, 11)]), (36, [('A1', 37, 37), ('A0', 11, 11)]), (41, [('A1', 42, 42), ('A0', 11, 11)])]
事件要素列表构建: eleList	[[['TMP', 'A0', 'Act', 'A1', 'QTY'], ['2020 年 12 月 31 日', '上海市', '新增', '境外输入性病例', '5']], [['TMP', 'A0', 'Act', 'A1', 'QTY'], ['2020 年 12 月 31 日', '上海市', '新增', '治愈出院病例', '7']], [['A0', 'Act', 'A1'], ['治愈出院病例', '来自', '俄罗斯']], [['A0', 'Act', 'A1'], ['治愈出院病例', '来自', '巴基斯坦']], [['A0', 'Act', 'A1'], ['治愈出院病例', '来自', '加拿大']], [['A0', 'Act', 'A1'], ['治愈出院病例', '来自', '安哥拉']], [['A0', 'Act', 'A1'], ['治愈出院病例', '来自', '西班牙']], [['A0', 'Act', 'A1'], ['治愈出院病例', '来自', '斯洛文尼亚']]]

在表 1 中,Ns 是通过 LTP 命名实体识别提取的命名实体集,一般包括人物、地名等,由于笔者主要研究特定区域的疫情动态,疫情区域被作为疫情主体,因此,命名实体集也是疫情主体集,代表各区域疫情发生或防疫的主体。

Pos 是通过对分词文本进行词性分析得到的初始事件要素集,作为事件要素列表的参照。初始事件要素主要是根据主谓宾三元组结构和词性中的动名词形式对分词块进行重新组合,提取的事件关键信息。

Srl 为语义角色标识集,是采用 LTP 工具包识别的以文本谓词为中心的各成分间的关系。其中,TMP 代表时间,A0 为语义角色中的主体标识,代表事件的施事者或者触发者,A1 为语义角色标注中的受事者。Act 为针对疫情信息的核心谓词,即疫情事件的触发动作,如“新增”“来自”等,围绕核心谓词,可以确立相关主体与对象之间的对应关系,为事件知识表示提供依据。按照表 1 的文本分析过程,以上海市卫生健康委员会的疫情播报信息为数据源,发生时间、疫情主体、防疫对象、疫情动态及状态等事件要素被确定。疫情事件要素构成了疫情事件本体的基本结构,为事件本体模型的设计奠定基础。为了便于本体的拓展和重用,本文进一步拓展了防疫资源要素,针对防疫主体所属资源进行扩充,如表 2 所示:

表 2　疫情事件关键要素

疫情主体	防疫对象	发生时间	触发动作	疫情状态	防疫资源
上海市	境外输入性病例	12 月 31 日	新增	新增境外输入性病例 5	医疗机构
西班牙	治愈出院病例	12 月 30 日	来自		
菲律宾	本地病例	12 月 29 日	……	新增治愈出院病例 7	
……	……	……			

通过事件要素识别的疫情事件的基本结构为疫情事件本体的层次结构及要素之间的基本关系确定提供直接依据，是疫情知识框架的基本雏形。基于疫情事件的关键要素，本文结合事件的四维度和六元组模型[16,31]，对疫情事件进行界定，通过五元组描述疫情事件，如式(1)：

$$E=(TIM,Sub,Act,Obj,Sta) \quad \text{式(1)}$$

在疫情事件五元组模型中，E 代表疫情事件，TIM 为疫情事件发生时间，Sub 为事件主体，主要是疫情事件中的施事主体和受事主体，包括疫情爆发区域、防疫区域等。Act 为事件动作或触发方式，如“新增”“现有”“累计”等。Obj 以防疫对象为主，即疫情事件的客体，包括染疫对象、潜在疫情风险群体等，也存在部分作为事件受事者的疫情主体等。Sta 为疫情事件状态，体现当前疫情的发展状况，笔者通过提取不同事件的状况作为其子集，如“新增治愈出院病例 5”。每个疫情事件实例归属于疫情事件类。

2.2　疫情事件本体模型

基于确定的事件要素，笔者构建了疫情事件本体模型，以刻画事件本体的基本结构，并为疫情事件知识的丰富与扩充奠定基础。疫情事件本体模型由疫情主体、防疫对象等概念，以及刻画概念间关系和实例状态的一系列属性组成，是事件知识表示的基本框架。基于疫情事件的特征，结合疫情事件关键要素和疫情事件本体应用的需要，笔者按照知识组织、丰富与更新、应用的思路，将疫情事件本体模型划分为概念和关系层、实例层和应用层 3 个层次，为疫情事件本体的构建和自动更新奠定基础。疫情事件本体模型如图 1 所示，其中蓝色线部分为根据文本内容自动填充的信息。

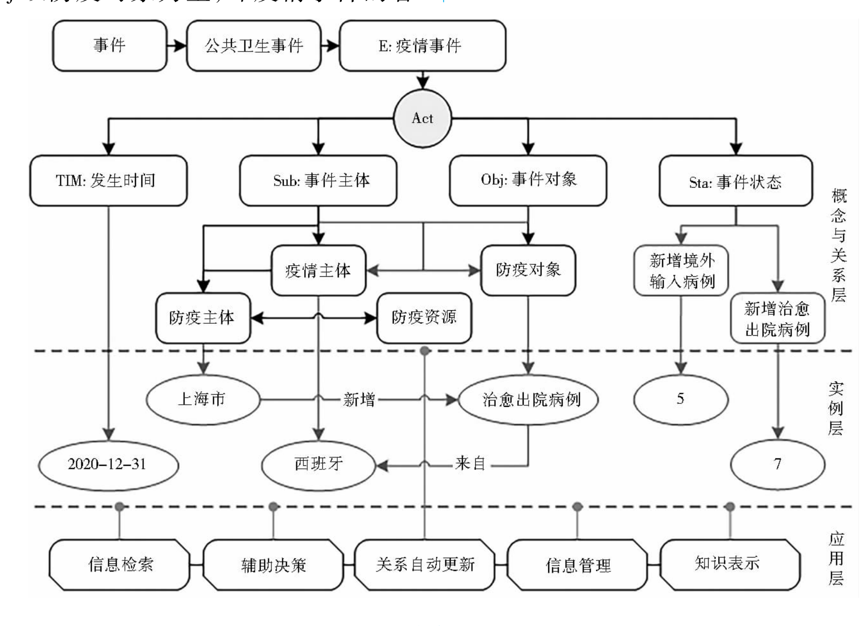


图 1　疫情事件本体模型

在疫情事件本体模型中，由事件触发动作引申出具体的事件动态(包括关系和属性)。因此，概念和关系层包含事件要素类及要素间的关系、要素的属性等，它们构成了事件本体的基本结构。基于概念与关系层，实例层包含由具体动作触发的具体疫情事件及事件相关的关系和状态，需要疫情数据和信息的填充与实例化。应用层基于概念层和实例层，为相关应用提供信息检索、属性和关系的动态维护、辅助决策、知识表示等功能。基于事件本体模型，事件本体涉及的基本概念、关系和属性展示在表 3 中。除此之外，还存在一些在本体自动填充过程中生成的概念、关系和属性。

表 3　疫情事件本体模型包含的基本概念和属性

名称	类别	定义域	值域	说明
疫情事件 E	类	–	–	疫情事件的集合
疫情主体 ES	类	–	–	疫情主体的集合
防疫主体 AES	类	–	–	防疫主体集合，ES 的子集
防疫对象 AEO	类	–	–	防疫对象的集合
防疫资源	类	–	–	防疫资的集合
医疗机构	类	–	–	防疫资的子集
涉及主体 rsub	对象属性	疫情事件	疫情主体	E 与 ES 间的关系
发生时间 tim	数据属性	疫情事件	文本值	疫情事件发生时间
疫情状态 sta	数据属性	疫情事件	文本值	疫情事件的状态
拥有资源	对象属性	防疫主体	防疫资源	AES 与防疫资源的关系

通过对疫情事件基本概念和关系的梳理，笔者进一步对疫情事件本体的层次关系进行初步搭建，构建疫情事件本体 epi，形成本体文件 epieve. owl，形成疫情知识库的基本架构，为疫情事件实例的自动填充和本体的重用与扩充做好准备。

3 疫情知识库的构建

本研究通过网络爬虫和文本分析技术采集并提取事件要素，构建事件要素列表，结合疫情事件特征设计事件本体自动构建与填充算法，将事件要素列表自动充实到事件本体中，并根据应用需求，不断融合开放知识图谱和领域本体，形成疫情知识库，为辅助决策提供支撑。笔者采用图 2 所示的事件本体驱动的疫情知识库构建过程，方案的实现分为疫情信息采集和预处理、疫情事件要素提取、疫情知识库构建与辅助决策 3 个主要环节，以上海市 2020 年的新冠肺炎疫情为例，将上海市卫健委播报的疫情动态作为核心数据源，对笔者所提出的策略进行应用和检验，构建核心疫情事件本体库。在此基础上，引入中文开放知识图谱数据，不断对疫情知识库进行扩充。

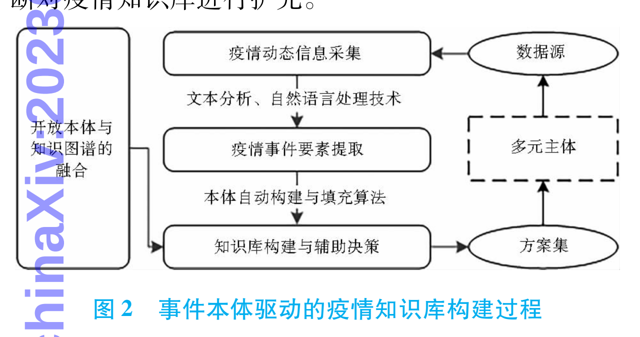


图 2 事件本体驱动的疫情知识库构建过程

3.1 疫情事件要素提取

在 2.1 节中，为了介绍疫情事件本体模型，笔者简要说明了通过 LTP 进行疫情文本分词、命名实体识别、语义角色标注，进而构建事件要素列表 eleList 的过程（见表 1），为了明确疫情事件的基本结构，本节将对疫情事件要素提取的细节做进一步说明。疫情事件要素的提取关键在于对事件的语义角色标注进行检验，补充缺失成分，并将语义角色标识与文本内容自动一一对应，以要素列表的形式存储。笔者设计了疫情事件要素列表构建算法（见算法 1），以准确描述客观事件。

按照 LTP 语义角色类型的约定，A0 代表事件触发者或施事者，A1 为受事者，QTY 代表数量，TMP 是事件发生时间，此外，语义角色列表中的首位元素代表根节点，即中心谓词，体现事件的动作，笔者以“Act”表示。对于事件主体缺失的情况，如“其中来自西班牙 1 例”，其语义角色标识为（36，[（'A1'，37，37）]），事件主体 A0 的部分丢失，笔者通过式（2）的前向遍历规则找到最近的事件主体，作为该三元组的主语，即“治愈出院病例来自西班牙”。其中，“ltp”为中文自然语言处理工具 LTP，“ltp. seg”和“ltp. srl”分别对疫情文本进行分词和语义角色标注处理，处理产生的数据集参照表 1 的示例。

$$A0'_t = segs(s)$$

$$s.t.\begin{cases} s = Srl_{i-1}^{1,n}(k) \\ n = len(Srl_{i=1}^{1}(k)) \\ n = n \ \ if \ k \in \{A0, A1\} \\ n = n-1 \ \ if \ k \notin \{A0, A1\} \end{cases} \qquad 式(2)$$

算法 1 疫情事件要素列表构建算法

```
输入：epiText      # epiText 为疫情文本
输出：eleList      # eleList 为疫情事件要素列表
Segs = [ltp.seg(epiText)] # 文本分词，Segs 为分词集
Srl = [ltp.srl(epiText)] # 语义角色标注，Srl 为语义角色集
eleList = [], ek = [ ], ev = [ ]
for srl in Srl and segs in Segs do
    eve = [ ] # 初始化事件片段
    ek.append('Act'), ev.append(segs[srl[0]]) # 事件触发动作
    for s in srl[1] do
        t = '', ek.append(s[0])
        for i in range(s[1], s[2] + 1) do t += segs[i] endfor
        ev.append(t)
        eve = [ek, ev]
    endfor
    if 'A0' not in eve[ek] then
        eve[ek].append('A0'), eve[ev].append(A0'_i←eq.(2))
    endif
    eleList.append(eve)
endfor
return eleList
```

基于自然语言处理的事件要素提取策略适合于事件概要描述层级的知识元素的提取，其中主要涉及词性分析和语义角色标注，并被广泛应用于文献[31，35]等事件要素分析的研究中。为了强化本文疫情事件要素提取策略的有效性和针对性，区分防疫主体 AES、疫情主体 ES 和防疫对象 EO 等概念并与事件要素列表对应，命名实体匹配、基于词性的初始事件要素对照等机制被引入，将事件要素与事件基本概念对应（见图 3），具体的实现逻辑体现在算法 2 疫情事件本体的自动构建环节。

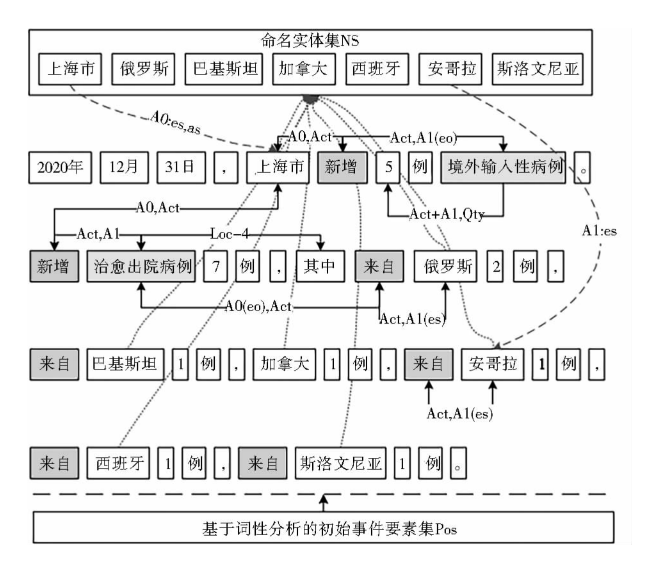


图 3　疫情事件要素与事件概念的对应

3.2　疫情知识库自动构建

基于提取的疫情事件要素，笔者进一步设计了算法 2，将相关要素转换为三元组形式，并自动创建，生成三元组图结构语义本体数据，形成疫情知识库的基本框架。基于疫情事件本体模型，疫情知识库自动构建算法对相关概念、对象属性和数据属性进行区分，自动构建并完善本体框架，然后将事件要素区分为实例、关系、属性及属性值，以三元组形式依次自动填充到本体架构中。

针对事件主体与对象的识别与区分，首先通过 LTP 对事件文本的所有命名实体进行识别并提取，提出命名实体匹配机制。通过分析发现，疫情事件中的命名实体全部为全球区域，即为疫情事件中的事件主体。命名实体匹配机制是将被标识为 A0 或 A1 的事件主体与被识别的命名实体集 Ns 的元素进行匹配，匹配到的 A0 元素为防疫主体 as，匹配到的 A1 元素为疫情主体 es，未能匹配到的元素是防疫对象 eo 或其他，相关概念由此被一一对应。

算法 2 首先将疫情事件要素与事件概念对应，通过三元组形式构建概念、实例间的关系和实例的属性，基于唯一资源标识 URI 不重复地将类、属性添加到事件本体，由此构建本体的层次结构和基本关系。在此基础上，通过事件要素列表读取疫情事件要素，构建疫情事件实例，对实例要素进行判断，若要素对应于事件主体 A0、时间 TMP、受事者 A1、量词 QTY 等类型，则构建要素的 URI 标识；其次，判断本体中是否存在相应的成分，若不存在，将相关要素自动添加到本体，并填充要素之间的关系和属性值。最后将填充后的本体数据返回，完成疫情事件本体的自动构建。基于 Protégé 平台可视化的疫情事件和防疫主体实例见图 4 和图 5。

算法 2 疫情事件本体自动构建与填充算法

```
输入:eleList, epi, Ns, Pos #事件要素列表,事件本体,命名实体集,基于词性的要素集
输出:epi
Class(E,ES,EO,AES) ∈ epi, AES ⊆ ES
Property(rsub,tim,sta) ∈ epi
for eve[ele,seg] ∈ eleList and eve.seg ∈ segs do   #seg 为对应的句子成分
    as = es = eo = act = act1 = sta = time = '', num = 0
    ek = eve.ele, ev = eve.seg
    if ek == A0 or ek == A1 and ev in Ns and ev in Pos(as,eo,so) then
        if ek == A0 then as = ev, epi.add((as, type, AES))
        else es = ev, epi.add((es, type, ES))
        endif
    elif ek == A1 and ev not in Ns and ev in Pos(as,eo,so) then
        eo = ev, epi.add((eo, type, EO))
    elif ek == Act or ek == QTY or ek == TMP then
        if ek == Act then act = ev, epi.add((act, type, objectProperty))
        elif ek == QTY then num = ev
        else time = ev
        endif
    endif
    e = 'EE' + time, epi.add((e, type, E)) # e 为疫情事件
    act1 = act + eo, epi.add((act1, type, datatypeProperty))
    epi.add((act1, subPropertyOf, sta))
    epi.add((e, tim, time))
    epi.add((e, rsub, es))
    epi.add((as, act, eo))
    epi.add((e, act1, num))
endfor
return epi
```



图 4　事件本体中的疫情事件实例



图 5　事件本体中的防疫主体实例

通过算法2,被逐步充实的疫情事件实例如图6所示。通过算法的循环运行,新的疫情信息得以不断补充,可以实现疫情知识库的自动更新,疫情知识逐渐丰富。

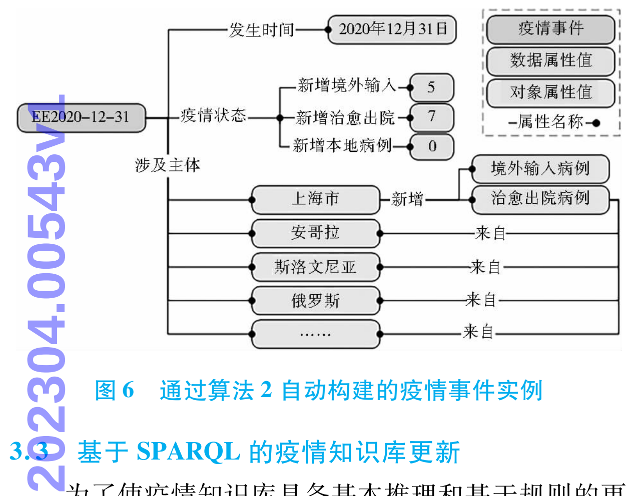


图 6　通过算法 2 自动构建的疫情事件实例

3.3　基于 SPARQL 的疫情知识库更新

为了使疫情知识库具备基本推理和基于规则的更新能力,以支持辅助决策,笔者设计了基于查询语言 SPARQL 的知识更新策略。一方面,疫情知识库可以结合算法 1 和算法 2 自动提取事件要素并进行自动更新;另一方面,也需要结合特定的应用需要,设计个性化的规则来丰富知识库。笔者从高风险事件和高风险主体两个角度设计基于 SPARQL 的查询规则,并结合算法 3,实现知识的更新。具体的查询规则如表 4 所示:

表 4　基于 SPARQL 的知识库信息更新规则

规则编号	基于 SPARQL 的对象提取	功能说明
Q1	SELECT ? e ? n WHERE { ? e epi:新增境外输入性病例 ? n. ? e epi:发生时间 ? t. FILTER(? n >X && ? t > starttime)}	将自时间 starttime 以来新增境外输入病例数量大于 X 例的事件提取,识别为高风险事件
Q2	SELECT ? es (count(? e) as ? en) WHERE{ ? e epi:涉及主体 ? es. FILTER NOT EXISTS { ? es rdf:type epi:防疫主体}} GROUP BY ? es HAVING(? en > Y)	将涉及事件数超过 Y 的疫情主体提取,作为高风险主体

算法 3 基于 SPARQL 的疫情知识库更新算法
输入: epi, X, starttime, Y
输出: epi

```
Hre = epi.query(Q1, X, starttime) #执行 SPARQL 查询
Hrs = epi.query(Q2, Y)
if Hre ! = [ ] then
    epi.add((HRE, subClassOf, E)) #创建“高风险事件”类 HRE
    for e in Hre do epi.add((e, type, HRE)) endfor
endif
if Hrs ! = [ ] then
    epi.add((HRS, subClassOf, ES)) #创建“高风险主体”类 HRE
    epi.add((ren, type, datatypeProperty)) #创建“涉及事件数”数据属性
    for (es,en) in Hrs do
        epi.add((es, type, HRS)) # es 为高风险主体
        epi.add((es, ren, en)) # en 为涉及事件数量
    endfor
endif
return epi
```

算法 3 按照 **Q1**、**Q2** 的查询规则提取高风险对象,创建分别归属于疫情事件和疫情主体的两个子类,将两类高风险对象分别归类,并针对后面新添加的信息进行自动更新操作,以支持在风险识别下的应急辅助决策。为了检验算法,笔者分别设置了相应的查询参数,具体结果如表 5 所示:

表 5　基于 SPARQL 的疫情知识库信息更新示例

查询参数	参数值	更新类	更新实例	更新属性
X	10	高风险事件	EE2020-12-19 EE2020-12-22	无
starttime	2020-12-25T23:59:59			
Y	40	高风险主体	俄罗斯 英国 菲律宾 美国	涉及事件数

3.4　结果分析

结合大数据背景下突发公共卫生事件信息管理和应急决策智能化的趋势,笔者以新冠肺炎疫情为例,对疫情知识表示策略展开研究。在数据源方面,选取上海市卫健委的疫情播报信息为基本数据源,采用网络爬虫动态爬取自 2020 年 1 月以来的疫情信息,截至 12 月 31 日,共提取 320 条语料,构建疫情语料库。在方法工具方面,运用 Python 语言对文本进行规范化,采用哈尔滨工业大学的语言技术平台工具包 LTP 进行语料的分词、命名实体识别、词性分析和语义角色标注,通过算法 1 对语义角色标识和语料进行匹配,构建事件

熊励，王成文，王锟. 基于事件本体的疫情知识库构建策略[J]. 图书情报工作,2021,65(14):138－148.

要素列表 eleList;运用算法 2 将要素列表 eleList 中的元素填充到疫情事件本体 epi,同时实现疫情实例相关信息的自动更新。笔者将疫情事件本体作为疫情知识库的载体,借助 SPARQL 实现算法 3 的查询规则,完成疫情知识库的动态更新。通过文本实验,最终获取包含 6 个类、4 个对象属性、9 个数据属性、410 个实例、1 121 对关系和 918 条数据属性值的疫情领域本体,初步实现了一个面向疫情的事件本体知识库,为疫情知识的组织与共享提供了参考和借鉴。

为了进行知识拓展,笔者分别选取了来自网络(https://yyk.99.com.cn/)的防疫主体所属的防疫资源数据(数据集 1)、中文开放知识图谱平台(http://www.openkg.cn/)"新冠开放知识图谱.事件"(数据集 2)和"新冠开放知识图谱.流行病"(数据集 3),对疫情知识库进行丰富和扩充。防疫资源是防疫主体(上海市)的三甲医院,其数据通过网络爬虫爬取,并采用算法 2 的逻辑,按照"防疫资源→医疗机构→医院"的层次来扩充,示例如图 7 所示:



图 7　对疫情知识库扩充的防疫资源示例

数据集 2 和数据集 3 分别对应于事件模式和流行病的 xlsx、json 格式数据(见图 8),笔者运用 python 语言对其进行解析,将 id、type、label、domain、range、subClassOf、subProperty 等关键字段转换为概念、示例和属性,基于算法 2 的思路分别构建了 event 和 covid 两个辅助本体,并借助 protégé 本体编辑工具将这两个本体合并到疫情事件本体中,最终的疫情知识库基本结构如图 9 所示:

不安全 | openkg1.oss-cn-beijing.aliyuncs.com/05e9af15

aph":[
{
"@id":"http://www.openkg.cn/2019-nCoV/event/class/C0",
"@type":"http://www.w3.org/2000/01/rdf-schema#Class",
"label":{
"@language":"zh",
"@value":"事件"
}
},
{
"@id":"http://www.openkg.cn/2019-nCoV/event/class/C1",
"@type":"http://www.w3.org/2000/01/rdf-schema#Class",
"label":{
"@language":"zh",
"@value":"公共卫生事件"
},
"subClassOf":"http://www.openkg.cn/2019-nCoV/event/class/C0"
},

(a) 数据集2

不安全 | openkg1.oss-cn-beijing.aliyuncs.com/56e2

{
"@id" : "http://www.openkg.cn/COVID-19/epidemiology/class/C61",
"@type" : "rdfs:Class",
"rdfs:label": "流行病",
"rdfs:subClassOf": {
"@id": "http://www.openkg.cn/COVID-19/epidemiology/class/C6"
}
},
{
"@id" : "http://www.openkg.cn/COVID-19/epidemiology/class/C7",
"@type" : "rdfs:Class",
"rdfs:label": "流行病学调查"
},
{
"@id" : "http://www.openkg.cn/COVID-19/epidemiology/class/C8",
"@type" : "rdfs:Class",
"rdfs:label": "疫源地"
},

(b) 数据集3

图 8　数据集 2 和 3 的数据片段

通过 3 个数据集的扩充,疫情知识得到了扩展,说明基于多源数据的事件本体知识拓展策略的可行性。这一方案为将疫情事件与资源、地理、人群、疾病、政策等领域知识的融合提供了参考和借鉴,为基于多源数据的应急决策提供了可借鉴的方案,也为进一步的疫情知识语义推理和重用奠定了基础。基于事件本体的疫情知识库自动构建结果表明,结合命名实体识别、词性分析和语义角色标注的自然语言处理等技术的本体自动构建与填充策略能够实现从数据采集、数据预处理、事件要素提取、事件本体自动构建和填充、知识库自动更新与拓展的疫情信息管理与知识表示,提升了知识库构建的效率并降低了人工成本。但是,疫情事件本体自动构建策略较大程度上依赖于文本预处理与自然语言处理策略,需要开发对应的算法来驱动本体的自动化构建。

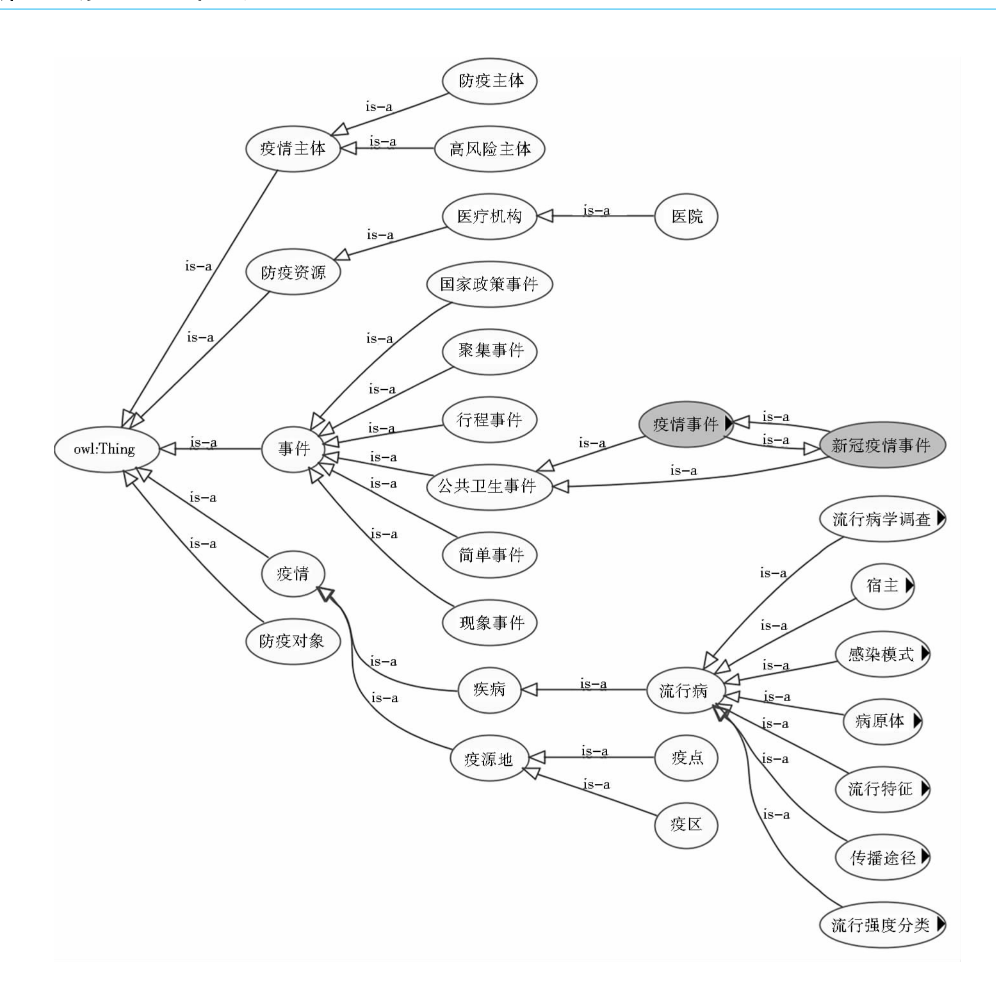


图 9　疫情事件本体扩充后的概念结构

4 总结

动态、智能的信息管理与知识表示是疫情应急管理与智能决策的迫切需求,也是大数据时代应急管理面临的一项挑战。本文的研究结果表明,综合网络爬虫、自然语言处理和事件本体的策略能够实现疫情信息的自动采集、事件要素提取、知识库自动构建和更新,进而为疫情信息管理和知识表示的智能化创造条件。在疫情知识的扩展和重用方面,多源数据可以被解析为概念和实例关系类型,并以三元组形式填充到事件本体,或被融合到其他的知识架构中。笔者所提出的本体自动构建策略为疫情知识表示和动态更新提供了参考和借鉴。

在未来的研究中,可以选取更加全面和权威的多源数据,并设计适用性和可扩展性更强的自然语言处理方法,以实现精确可靠的事件要素提取。此外,在以事件为线索的动态知识表示方面,将围绕事件主题的大规模领域知识集成以实现知识的自动扩展,从而为智能应急决策提供支撑的研究将是一个有重要意义的方向。

参考文献:

[1] 曹振祥,储节旺,郭春侠. 面向重大疫情防控的应急情报保障体系理论框架构建——以 2019 新型冠状病毒肺炎疫情防控为例[J]. 图书情报工作, 2020,64(15):72-81.

[2] 李进华. 面向大数据时代的重大疫情信息管理理论框架及其应用[J]. 现代情报, 2020,40(7):25-33,51.

[3] NI Z J, RONG L L, WANG N, et al. Knowledge model for emergency response based on contingency planning system of China[J]. International journal of information management, 2019,46:10-22.

[4] 丁波涛. 人工智能时代的疫情信息管理:挑战与变革[J]. 图书情报知识, 2020(6):109-116.

[5] LEE W B, WANG Y, WANG W M, et al. An unstructured information management system (UIMS) for emergency management[J]. Expert systems with applications, 2012,39(17): 12743-12758.

[6] DORASAMY M, RAMAN M, KALIANNAN M. Integrated community emergency management and awareness system: a knowledge management system for disaster support[J]. Technological forecasting and social change, 2017,121: 139-167.

[7] 郭骅,屈芳,战培志. 城市应急管理情报平台构建研究[J]. 图书情报工作, 2018,62(6):93 – 104.

[8] 徐蓓蓓,杨子江,朱世伟,等. 基于本体的突发事件舆情知识库建设研究[J]. 情报杂志, 2019,38(4):132 – 137.

[9] 赵又霖,庞烁,吴宗大. 社会感知数据驱动下突发事件应急管理的时空语义模型构建研究[J]. 情报科学, 2021,39(2):44 – 53.

[10] 王芳,杨京,徐路路. 面向火灾应急管理的本体构建研究[J]. 情报学报, 2020,39(9):914 – 925.

[11] WU H T, ZHONG B T, MEDJDOUB B. An ontological metro accident case retrieval using CBR and NLP[J]. Applied sciences, 2020,10(15): 5298.

[12] 高珊,王文俊,杜磊,等. 传染病应急案例共享本体模型研究[J]. 计算机应用, 2010,30(11):2924 – 2927.

[13] 方安,洪娜,高东平,等. 传染病本体构建及其在知识服务平台中的应用[J]. 现代图书情报技术, 2012(1):7 – 12.

[14] HOGAN W R, WAGNER M M, BROCHHAUSEN M, et al. The Apollo structured vocabulary: an OWL2 ontology of phenomena in infectious disease epidemiology and population biology for use in epidemic simulation[J]. Journal of biomedical semantics, 2016, 7: 50.

[15] 陈晓慧,刘俊楠,徐立,等. COVID-19 病例活动知识图谱构建——以郑州市为例[J]. 武汉大学学报(信息科学版), 2020, 45(6):816 – 825.

[16] JOSHI A, KARIMI S, SPARKS R, et al. Survey of text-based epidemic intelligence: a computational linguistics perspective[J]. ACM computing surveys, 2019,52(6): 119.

[17] TUDORACHE T. Ontology engineering: current state, challenges, and future directions[J]. Semantic web, 2020, 11(1): 125 – 138.

[18] LV Z M, PENG R. A novel meta-matching approach for ontology alignment using grasshopper optimization[J]. Knowledge-based systems, 2020,201: 106050.

[19] ZHONG Z, LIU Z, LIU W, et al. Event ontology and its evaluation[J]. Journal of information and computational science, 2010,7(1): 95 – 101.

[20] LIU W, JIANG L, WU Y S, et al. Topic detection and tracking based on event ontology[J]. IEEE access, 2020, 8: 98044 – 98056.

[21] ZHONG Z M, LIU Z T, LI C H, et al. Event ontology reasoning based on event class influence factors[J]. International journal of machine learning and cybernetics, 2012,3(2): 133 – 139.

[22] GOY A, MAGRO D, ROVERA M. On the role of thematic roles in a historical event ontology[J]. Applied ontology, 2018,13(1): 19 – 39.

[23] 刘炜,丁宁,杨竣辉,等. 针对环境污染突发事件领域的事件本体模式[J]. 计算机科学与探索, 2016,10(4): 466 – 480.

[24] FERTIER A, BARTHE-DELANOE A M, MONTARNAL A, et al. A new emergency decision support system: the automatic interpretation and contextualisation of events to model a crisis situation in real-time[J]. Decision support systems, 2020,133: 113260.

[25] 朱宇倩,廖慧敏,罗小娟,等. 安全信息认知事故的事件本体模型研究[J]. 情报杂志, 2020,39(7): 151 – 158.

[26] SINGLE J I, SCHMIDT J, DENECKE J. Knowledge acquisition from chemical accident databases using an ontology-based method and natural language processing[J]. Safety science, 2020,129: 104747.

[27] 唐英英,刘炜,刘菲京,等. 基于事件本体的 Web 服务动态组合[J]. 华中科技大学学报(自然科学版), 2013,41(S2): 22 – 25.

[28] KUPTABUT S, NETISOPAKUL P. Event extraction using ontology directed semantic grammar[J]. Journal of information science and engineering, 2016,32(1): 79 – 96.

[29] 朱文跃,刘炜,刘宗田. 基于事件本体的新闻个性化推荐[J]. 计算机工程, 2019,45(6): 267 – 272, 279.

[30] 刘思含,刘旭红,刘秀磊. 事件本体表示模型与构建综述[J]. 北京信息科技大学学报(自然科学版), 2018,33(2): 35 – 40, 81.

[31] 朱文跃,刘宗田. 基于事件本体的突发事件领域知识建模[J]. 计算机工程与应用, 2018,54(21): 148 – 155.

[32] GURBUZ O, RABHI F, DEMIRORS O. Process ontology development using natural language processing: a multiple case study[J]. Business process management journal, 2019, 25(6): 1208 – 1227.

[33] REYES-ORTIZ J A. Criminal event ontology population and enrichment using patterns recognition from text[J]. International journal of pattern recognition and artificial intelligence, 2019,33(11): 1940014.

[34] 王思丽,祝忠明,刘巍,等. 基于深度学习的领域本体概念自动获取方法研究[J]. 情报理论与实践, 2020,43(3):145 – 152, 144.

[35] MAO Q, LI X, PENG H, et al. Event prediction based on evolutionary event ontology knowledge[J]. Future generation computer systems, 2021,115: 76 – 89.

作者贡献说明:

熊励:确定选题,提出研究思路,指导论文修改;

王成文:设计研究方案,进行实验和撰写论文;

王锟:算法校验与论文修改。

Construction Strategy of Epidemic Knowledge Base Based on Event Ontology

Xiong Li[1] Wang Chengwen[1] Wang Kun[1,2]
[1] School of Management, Shanghai University, Shanghai 200444
[2] Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney 2007

Abstract: [**Purpose/significance**] The fragmented and unstructured information of the epidemic brings challenges to emergency decision-making. To support the digitization of emergency decision – making and to promote intelligent emergency management, combining natural language processing and event ontology to realize the automation of epidemic information management and knowledge representation. [**Method/process**] An automatic construction strategy of domain ontology knowledge base based on web crawler, natural language processing, and event ontology was proposed. First, web crawlers and natural language processing were used for information collection and automatic extraction of event elements, and an epidemic event ontology model was built on this basis. Then, the algorithms for the ontology construction and update were designed, and the automatic construction and expansion for the event ontology was completed by them. [**Result/conclusion**] The results show that the proposed strategy has the feasibility of dynamic management and automatic update of epidemic information, and event ontology can describe events effectively and create conditions for knowledge expansion. This study also provides a reference for the research and practice of emergency decision-making.

Keywords: event ontology epidemic information and knowledge knowledge base